



On an extension of Krivovichev's complexity measures

Wolfgang Hornfeck*

Institute of Physics of the Academy of Sciences of the Czech Republic, Na Slovance 2, 182 21 Praha 8, Czech Republic.

*Correspondence e-mail: wolfgang.hornfeck@web.de, hornfeck@fzu.cz

Received 9 April 2020

Accepted 18 May 2020

Edited by A. Altomare, Institute of
Crystallography - CNR, Bari, Italy**Keywords:** Shannon entropy; Krivovichev
complexity; strong additivity; crystal structure
classification; structural complexity.

An extension is proposed of the Shannon entropy-based structural complexity measure introduced by Krivovichev, taking into account the geometric coordinational degrees of freedom a crystal structure has. This allows a discrimination to be made between crystal structures which share the same number of atoms in their reduced cells, yet differ in the number of their free parameters with respect to their fractional atomic coordinates. The strong additivity property of the Shannon entropy is used to shed light on the complexity measure of Krivovichev and how it gains complexity contributions due to single Wyckoff positions. Using the same property allows for combining the proposed coordinational complexity measure with Krivovichev's combinatorial one to give a unique quantitative descriptor of a crystal structure's configurational complexity. An additional contribution of chemical degrees of freedom is discussed, yielding an even more refined scheme of complexity measures which can be obtained from a crystal structure's description: the six C's of complexity.

1. Introduction

In a series of recent articles, Krivovichev (2012*a,b*, 2013*a,b*, 2014*a,b*, 2016, 2017; Krivovichev *et al.*, 2017, 2018; Krivovichev & Krivovichev, 2020) proposed an elegant way of measuring the structural and topological complexity of crystal structures in terms of their Shannon information amount, or Shannon entropy.

1.1. Shannon entropy

Shannon (1948*a,b*) quantified the information amount H encoded in a message constituted of P entities falling into N equivalence classes as (in units of bits)

$$H_N(\mathcal{P}) = H_N(p_1, p_2, \dots, p_N) = \sum_{i=1}^N L(p_i), \quad (1)$$

in which

$$L(x) = \begin{cases} 0 & \text{for } x = 0 \text{ or } x \text{ non-definite,} \\ -x \log_2 x & \text{for } x > 0, \end{cases} \quad (2)$$

and where

$$p_i = P_i/P \quad (3)$$

denotes the probability of occurrence of the i th symbol, which is calculated as the quotient of the number of the i th symbol, P_i , to the total number of symbols,

$$P = \sum_{i=1}^N P_i. \quad (4)$$

$$\text{crystallographic} \quad (Z, M, A) = \underbrace{\text{chemical} \quad (Z)}_{\text{compositional} \quad (Z, M)} + \underbrace{\text{combinatorial} \quad (M) + \text{coordinational} \quad (A)}_{\text{configurational} \quad (M, A)}$$

The N individual probabilities p_i are the elements of a discrete finite probability distribution,

$$\mathcal{P} = [p_1, p_2, \dots, p_N] = [p_i]_{i=1}^N = [P_i]_{i=1}^N / P, \quad (5)$$

where $N = |\mathcal{P}|$. The continuous case is also possible, as well as various other generalizations (Aczél & Daróczy, 1975), although these are of no further interest in the context of this work. Note that the probabilities may form the mathematical object of a multiset, meaning that, in contrast to a set, multiple instances of any element are allowed. Permuting their order leaves the Shannon entropy invariant (symmetry property), as does including/excluding probabilities $p_i = 0$ (expansibility property). Note also that the following two conditions hold:

$$\begin{aligned} \text{(i)} \quad & 0 \leq p_i \leq 1, \\ \text{(ii)} \quad & \sum_{i=1}^N p_i = 1. \end{aligned} \quad (6)$$

The second criterion guarantees the completeness of the probability distribution (a generalization of the Shannon entropy due to Rényi allows for incomplete probability distributions too; see Aczél & Daróczy, 1975, pp. 26–27). For later use and reasons of clarity we also define a simplified version of equation (1) using a slightly different notation,

$$H_N(\mathcal{P} \times \mathcal{P}) = H(P_1, P_2, \dots, P_N), \quad (7)$$

which just lists the integer enumerators P_i , since N and P can be easily inferred from the number of enumerators and their sum, respectively. In the case of equidistributed probabilities, with the number of equivalence classes being a maximum, $N = P$, and with all enumerators being unity, $P_1 = P_2 = \dots = P_P = 1$, we further condense the entropy symbol to

$$H_P(\mathcal{P} \times \mathcal{P}) = H(P_1, P_2, \dots, P_P) = H(1, 1, \dots, 1) = H_P. \quad (8)$$

An axiomatic treatment of information theory has shown that the Shannon entropy really is the most natural measure of information with respect to a number of intuitive, plausible and desirable mathematical properties expected from such a measure (Aczél & Daróczy, 1975).

1.2. Krivovichev complexity

Shannon's (1948*a,b*) entropy formula was applied to chemical graphs quite early on by Rashevsky (1955), while later Bertz (1981, 1983) based a variety of molecular complexity measures upon it. Its crystallographic application due to Krivovichev is based on interpreting a crystal structure as a message consisting of atoms. The subdivision of M atoms into N equivalence classes is given by the distinct types of crystallographic orbits occurring in a crystal structure, each one encompassing a set of symmetrically equivalent atoms. These crystallographic orbits are represented by the Wyckoff positions associated with a given space-group type.

Note that in order to maintain a consistent way of assigning a complexity value to a crystal structure, the number of atoms refers to the reduced unit cell of the crystal, *i.e.* the unique (apart from orientation) primitive unit cell, which fulfils certain algebraic conditions imposed on the basis vectors of

the lattice. For centred non-primitive unit cells the number of atoms is larger by a factor of 2, 3 or 4 depending on the centring type.

Now, in its crystallographic interpretation, the probability of occurrence is given as the quotient $m_i = M_i/M$ of the multiplicities M_i of occupied Wyckoff positions, with the total number of atoms M given as their sum [*cf.* equation (4)] and the individual probabilities forming the elements of a probability distribution \mathcal{M} [*cf.* equation (5)].

From this, and following equation (1), the conventional crystallographic Shannon entropy is defined as

$$I_G = H_N(\mathcal{M}) = \sum_{i=1}^N L(M_i/M), \quad (9)$$

measuring a crystal structure's complexity in bits per atom. Here, *i.e.* in Krivovichev's notation, the index G originates from 'graph', since the same formula can be used to measure the information content of an abstract mathematical graph. We will stick to this notation for the moment, in order to maintain cross-referencing with the existing literature. However, we will change the notation later according to $I_G = I_M$, in order to maintain a general scheme for all yet-to-be-introduced univariate Shannon entropies.

A number of derived complexity measures follow. First, the maximal information content of a crystal structure, also measured in bits per atom, and given as

$$I_{G,\max} = H_M = \log_2 M. \quad (10)$$

Assuming fully symmetrically independent atoms, with the number of equivalence classes matching the number of atoms ($N = M$), it represents the case of equidistributed probabilities, $m_i = 1/M$. In information theory contexts this maximal Shannon entropy is sometimes also known under the name of Hartley entropy. Second, the normal information content of a crystal structure,

$$I_{G,\text{norm}} = \frac{I_G}{I_{G,\max}}, \quad (11)$$

represents a dimensionless quantity ranging between zero (all atoms are symmetrically equivalent, $N = 1$, $M_1 = M$) and unity (no atoms are symmetrically equivalent, $N = M$, $M_i = 1$). Third, the total information content of a crystal structure,

$$I_{G,\text{total}} = M \cdot I_G = M \cdot I_{G,\max} + \sum_{i=1}^N L(M_i), \quad (12)$$

takes into account the size of the system, measured in bits per unit cell. (For the derivation of the rightmost equation from the middle one, see Appendix A.)

It is noteworthy that all these measures are independent of the crystal structure's metrics, being combinatorial in nature instead, though it is also possible to define an information density, $\rho_{\text{inf}} = I_{G,\text{total}}/V_{\text{red}}$, measured in bits per cubic ångström, taking into account the volume V_{red} of the reduced unit cell.

Krivovichev applied these information-theory-based complexity measures to a wide range of compounds, in

particular inorganic crystal structures (Krivovichev; 2012*a,b*, 2014*a,b*), zeolites and other minerals (Krivovichev, 2013*a,b*; Krivovichev, 2017; Krivovichev *et al.*, 2018), combining these ideas on structural complexity with the notion of algorithmic complexity proposed by Kolmogorov and the concept of crystal-structure generation by cellular automata (Krivovichev, 2014*a*), as well as relating them to the symmetry of minerals (Krivovichev & Krivovichev, 2020). Furthermore, relations between the structural complexity of crystals and their thermodynamic properties were explored, in particular regarding the configurational entropy of a crystal structure (Krivovichev, 2016) and the enthalpy-governed crystallization sequence of polymorphs following the Ostwald step rule (Krivovichev *et al.*, 2017), making this approach firmly interconnected with energy principles.

1.3. Combinatorial foundations

In order to get a feeling for the Krivovichev complexities of crystal structures in an abstract and most general sense, namely by generating a spectrum of potential complexity values in a systematic way, one can study the sequence of normalized entropies,

$$H_{N,\text{norm}}(i) = \frac{1}{\log_2 N} \sum_{j=1}^{|\mathcal{P}_N(i)|} L\left(\frac{\mathcal{P}_N(i,j)}{N}\right), \quad (13)$$

in which $\mathcal{P}_N(i, j)$ denotes the j th summand (part) of the i th member $\mathcal{P}_N(i)$ of the set \mathcal{P}_N of canonically ordered integer partitions for the number N . Here, $|\mathcal{P}_N(i)|$ represents the length of a single integer partition taken from this set. The total number of integer partitions for a given non-negative integer N is determined by the number-theoretical partition function $p(N) = |\mathcal{P}_N|$.

Integer partitions arise in this setting because of the subdivision of all M of a crystal structure's atoms into N crystallographic orbits of symmetry-equivalent ones, with the integer multiplicities of the associated Wyckoff positions always summing up to the total number of atoms, $M = M_1 + M_2 + \dots + M_N$, thereby forming an integer partition $[M_1, M_2, \dots, M_N]$. Thus, any crystal structure can always be associated with a corresponding integer partition. Any integer partition, by application of equation (5), then yields a probability distribution from which, by application of equation (1), its Shannon entropy can be calculated.

Take, for instance, the set of integer partitions for $N = 4$,

$$\mathcal{P}_4 = \{[4], [3, 1], [2, 2], [2, 1, 1], [1, 1, 1, 1]\}, \quad (14)$$

with $p(4) = |\mathcal{P}_4| = 5$, which yields

$$\begin{aligned} H_{4,\text{norm}}(1) &= L(4/4)/2 = 0.000, \\ H_{4,\text{norm}}(2) &= L(3/4)/2 + L(1/4)/2 = 0.406, \\ H_{4,\text{norm}}(3) &= L(2/4) = 0.500, \\ H_{4,\text{norm}}(4) &= L(2/4)/2 + L(1/4) = 0.750, \end{aligned} \quad (15)$$

$$\text{and } H_{4,\text{norm}}(5) = 2L(1/4) = 1.000,$$

respectively (all values in bits per freedom). As a general trend, and in agreement with intuition, it can be noted that the

Shannon entropy values increase with increasing differentiation of a partition (abstract crystal structure) and decrease with increasing integration (symmetry equivalence) in the opposite direction. Note, in particular, that the most condensed partition $\mathcal{P}_4(1) = [4]$ represents the case in which the normalized entropy measure is zero,

$$H_{N,\text{norm}}(1) = (\log_2 N)^{-1} H(N) = 0, \quad (16)$$

while the most expanded equidistributed partition $\mathcal{P}_4(5) = [1, 1, 1, 1]$ represents the case in which the normalized entropy measure becomes unity,

$$H_{N,\text{norm}}(N) = (\log_2 N)^{-1} H_N = 1, \quad (17)$$

thus defining natural complexity limits for any given partition.

An important thing to note in the approach of Shannon is that the size N of the partition is reflected only in the non-normalized entropy values H_N for the equidistributed case, since while

$$0 = H_1 < H_2 < H_3 < H_4 < \dots, \quad (18)$$

the non-normalized non-equidistributed entropy values $H(N)$ amount to zero independent of N :

$$0 = H(1) = H(2) = H(3) = H(4) = \dots \quad (19)$$

Thus, in the approach of Shannon, it is not the actual size of a system that is responsible for the complexity, *but how a system is organized into subsystems*, as highlighted by the stars-and-bars notation

$$\mathcal{P}_4(4) = [2, 1, 1] \Leftrightarrow \star\star | \star | \star \quad (20)$$

used in combinatorics.

Notably, this differs from an empirical definition of structural complexity, pragmatically based on the simple number of atoms in the reduced unit cell, such as used by Dshemuchadse & Steurer (2015) and Steurer & Dshemuchadse (2016) for the classification of complex metallic alloys. From among all compound classes scoring high on average in their structural complexity, intermetallics are infamous for harbouring some of the most complex crystal structures known, with $I_{G,\text{total}}$ commonly exceeding values of 10^3 bits per unit cell. The record-holding intermetallic compound $\text{Al}_{55.4}\text{Cu}_{5.4}\text{Ta}_{39.1}$, the most complex in terms of the simple number of atoms, 23 134, in its unit cell, exhibits a stunning Krivovichev complexity of $I_{G,\text{total}} = 48\,538.637$ bits per unit cell (Krivovichev, 2014*b*), thereby setting an upper limit on the scale of complexity values one has to expect.

In our attempt to understand the complex crystal structure of $\text{Na}_{11}\text{Hg}_{52}$ (Hornfeck & Hoch, 2015), with an already remarkable $I_{G,\text{total}} = 3936.056$ bits per unit cell (Tambornino *et al.*, 2015), we estimated its structural complexity in terms of counting, albeit in a rather *ad hoc* fashion, the compound's chemical and geometric degrees of freedom. In the remainder of this article we aim to refine the notions of chemical and geometric degrees of freedom, trying to make them precise and quantitative and extending Krivovichev's idea.

The rationale behind this approach is to continue stepping along a path trodden by Mackay (2001), after which the

complexity of a crystal structure can be defined by the number of parameters necessary to describe it, a notion similar to Kolmogorov's concept of algorithmic complexity, asking for the shortest possible set of rules necessary to describe a pattern.

At the outset, we motivate this pursuit with some questions: What about the geometric degrees of freedom? How can they be represented? Should they not be taken into account too? And preferentially in the same systematic scheme, for that matter?

2. Extending Krivovichev complexity: univariate case

Shannon's approach is very general: since it does not take into account the semantic content of a message, focusing on the relational structure existing between entities rather than their absolute meaning, it is widely applicable.

It is, however, in the specific application, when a meaning is established by interpreting what the entities under consideration are, and how their relational structure translates into a complete probability distribution, that the definition of the calculation of the Shannon entropy is based. It is this establishment of meaning which makes the application of the Shannon entropy a subjective choice, opening up the possibility of extending Krivovichev's complexity measures.

2.1. Univariate Shannon entropies

In order to exploit the advantages of a Shannon-type complexity measure in a more comprehensive way, Fig. 1 reviews the general framework for univariate entropy formulas.

The integer parameter X_i counting the size of an individual subsystem, out of N subsystems in total, together with the size X of the system defines a probability x_i . All probabilities x_i constitute a probability distribution \mathcal{X} . From this a Shannon entropy $H_N(\mathcal{X}) = I_X$ can be calculated, as well as its derived maximal, normal and total complexities after Krivovichev.

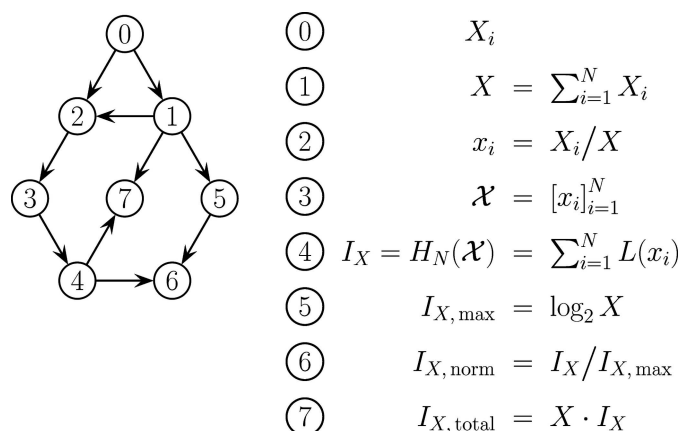


Figure 1

The general framework for univariate entropy formulas. The graph on the left illustrates the way in which the mathematical entities are derived from each other, eventually yielding the various discussed information measures.

While in the following special Shannon entropies will get their own symbol, in order to reflect their individual contributions, it seems best to address their units by the common phrase 'bits per freedom' where applicable, highlighting the intention to treat all degrees of freedom alike.

2.2. Crystal structures and Wyckoff sequences

Now, in order to extend the complexity measures of Krivovichev one has to identify other quantitative crystal structure descriptors (QCSDs; Mackay, 1984; Hornfeck, 2012) which lend themselves to a description using Shannon's and Krivovichev's approaches. In particular, the QCSDs should be of integer type, forming ratios with their totals in the denominator, thus facilitating the definition of a probability distribution and the corresponding entropy measures.

A complete geometric description of a crystal structure consists of three parts: (i) its symmetry, as given by the space-group type; (ii) its metrics, as given by the lattice parameters; and (iii) its atomic coordinates, as determined for an asymmetric unit, and given by the specification of the occupied Wyckoff positions.

For a given space-group type and setting, each Wyckoff position can be specified by a certain Wyckoff letter, where a to z and α define the possible alphabet (in its entirety only present for the space-group type $Pmmm$, No. 47). Thus, the geometry of a crystal structure can be represented very concisely by stating this information in a linear encoding, a crystal structure Wyckoff sequence, composed of the space-group type number (sometimes its Hermann–Mauguin symbol is used instead), followed by a list of Wyckoff letters, in reverse alphabetical order, each letter with its frequency of occurrence f_i assigned as a superscript (usually dropping $f_i = 1$).

When based on standardized crystal structure information, the Wyckoff sequence is a unique encoding of an abstract crystal structure. Only if additional crystallochemical classification criteria, such as distinguishable unit-cell axial ratios and atomic coordination environments, are taken into consideration does a finer classification into isopointal/isotypic structure types result (Parthé & Gelato, 1984; Parthé *et al.*, 1993).

In the following, structure type and Wyckoff sequence information were taken from *Pearson's Crystal Data Crystal Structure Database for Inorganic Compounds* (PCD; Villars & Cenzual, 2013) using standardized crystal structure data. Note that the number of distinct Wyckoff sequences is quite small, with only 15 503 cases represented in the aforementioned database.

While most of the parameters used in the description of crystal structures are numerical from the outset, two notable exceptions exist for most crystal structure descriptions in that the space-group-type symmetry and the atomic decoration of the Wyckoff sites are given by qualitative, not quantitative, descriptors only, namely by the Hermann–Mauguin symbol and the ones used for the chemical elements, respectively. This mixing of qualitative and quantitative statements is a major nuisance for any theoretical treatment that aims to be fully quantitative.

Table 1

Numerical parameters used in the description of crystal structures.

The definition ranges of the parameters are given by finite sets of integers, semi-open continuous intervals of real numbers and the positive real line, respectively.

| Parameter | Number set | Definition range |
|----------------------|--------------|--|
| Lattice parameters | \mathbb{R} | > 0 |
| Atomic numbers | \mathbb{Z} | $\{1, 2, \dots, 118\}$ |
| Atomic coordinates | \mathbb{R} | $[0, 1)$ |
| Atomic displacements | \mathbb{R} | > 0 |
| Site multiplicities | \mathbb{Z} | $\{1, 2, 3, 4, 6, 8, 12, 16, 24, 48\}$ |
| Site arities | \mathbb{Z} | $\{0, 1, 2, 3\}$ |
| Site occupancies | \mathbb{R} | $(0, 1]$ |

An easy workaround for the designation of the atom types is to use the atomic number Z_i of the chemical element instead, with the index i denoting the associated element of the probability distribution \mathcal{Z} . Note that, in the context of this work, the atomic number of a chemical element will always be represented by the indexed symbol Z_i , while the non-indexed symbol Z will be reserved for the sum total of all atomic numbers in the crystal structure, thus following the notation for univariate entropy measures as shown in Fig. 1. Note in particular that Z does *not* represent the number of formula units, contrary to the common use of the letter in the context of crystal structure description.

Table 1 gives a review of the numerical parameters involved in the description of crystal structures, together with the types of numbers and definition ranges for each parameter.

Note that we use the term ‘site arity’ as a shorthand for the degrees of freedoms of a given Wyckoff position with respect to its general atomic coordinates as tabulated in the *International Tables for Crystallography*, Vol. A (Aroyo, 2016). This notion follows the nomenclature in computer science, where arity is the number of arguments a function takes, e.g. $f(x, y)$ for a binary function. It is also reflected in the corresponding descriptors invariant, univariant, bivariate and trivariate, which describe the crystallographic cases of a fixed Wyckoff position or one with one, two or three degrees of freedom, respectively, following the nomenclature used for lattice complexes (Aroyo, 2016). In mathematical language, these are also known as constant, univariate, bivariate and trivariate, respectively. Finally, this also allows us to differentiate the distinct contributions to a combined Shannon entropy, where the general idea is to treat all degrees of freedom, say multiplicities *and* arities, on an equal footing.

2.3. Excursion I: an entropy for symmetry?

Still, the towering notion of symmetry appears to be absent from the listing given in Table 1. However, this is only partly true, since the symmetry is expressed by the splitting pattern of the multiplicities of the crystallographic orbits, and thus does not have to be accounted for separately.

However, since each space-group type is characterized not only by its group of lattice translations of infinite order, but also by its finite-order point group (crystal class) of rotation

and reflection operations, and because symmetry elements can be classified according to their group-theoretical order,

$$1 \rightarrow 1; \bar{1}, 2, \bar{2} (= m) \rightarrow 2; 3 \rightarrow 3; 4, \bar{4} \rightarrow 4; \bar{3}, 6, \bar{6} \rightarrow 6, \quad (21)$$

it becomes possible to calculate a Shannon entropy of a point group, in just the same way as described above, by accounting for all the orders of the individual symmetry elements present. This might be useful to establish an ordering of space-group types according to their entropies in those applications where a numerical ranking of objects according to their symmetry is needed. Indeed, in the context of pattern recognition the Shannon entropy has already been used to define an amount-of-symmetry detecting measure (Yodogawa, 1982).

2.4. Excursion II: on the choice of chemical degrees of freedom

In the previous section a choice is made to represent the chemical degrees of freedom of a crystal structure by the atomic number of its constituent elements. This choice is necessary in order to replace the qualitative information on the atom type with a quantitative parameter, for which the atomic number seems to be the most natural match. Yet is this choice meaningful? And if so, is it unique?

Regarding the first question, one might illustrate the idea with an example: do NaCl, KCl and RbCl, all of Wyckoff sequence 225, *ba*, differ in their chemical complexity? In terms of valence electrons, determining their chemistry, they do not. In terms of their total electron distributions, however, they do, with the electron counts for the ion pairs (M^+, X^-) being (10, 18) for NaCl, (18, 18) for KCl and (36, 18) for RbCl. Indeed, both K^+ and Cl^- in KCl share the same number of electrons, being isoelectronic to neutral argon, thereby veiling their difference in X-ray diffraction by their identical scattering contrast. Their corresponding entropies reflect this,

$$\begin{aligned} H(10, 18) &= 0.940 \quad (\text{NaCl}), \\ H(18, 18) &= 1.000 \quad (\text{KCl}), \\ H(36, 18) &= 0.918 \quad (\text{RbCl}), \end{aligned} \quad (22)$$

with the entropy being maximized for the equidistributed case. Thus, regarding the total electron distribution, RbCl is slightly less chemically complex than NaCl, which is itself slightly less chemically complex than KCl.

Concerning the second question, the answer is no, the choice is not unique, but rather a matter of definition. We mention only two alternatives. First, with respect to the concept of structure-type maps, another choice for differentiating the atom types could be the use of Mendeleev numbers (Pettifor, 1984, 1986). While these are integers too, they place the chemical elements in sequence according to a different logical order, thereby reflecting some of their properties in a better way than atomic numbers do. Second, one might also use the stoichiometric coefficients of the chemical formula associated with a crystal structure, as has been done by Siidra *et al.* (2014) and Krivovichev *et al.* (2018) for their $^{\text{chem}}I_G$ complexity measure.

2.5. Chemical, combinatorial and coordinational complexity

Thus, to summarize, there are just three integer crystal structure descriptors, the atomic numbers Z_i , the site multiplicities M_i and the site arities A_i , which can be used, in full analogy with Krivovichev's approach, for defining three corresponding sets (tetrads),

$$\mathcal{I}_X = \{I_X, I_{X, \max}, I_{X, \text{norm}}, I_{X, \text{total}}\}, \quad (23)$$

of four univariate information measures each, in which the subscript X denotes the type of attribute $X \in \{Z, M, A\}$. Note that, in this general scheme, the notation of the Krivovichev complexity measures [cf. equations (9) to (12)] changes according to $I_G \rightarrow I_M$. In addition, we choose to introduce another unifying notation, highlighting their nature as Shannon entropies,

$$\begin{aligned} H_{\text{chem}} &= I_Z = H_{|\mathcal{Z}|}(\mathcal{Z}), \\ H_{\text{comb}} &= I_M = H_{|\mathcal{M}|}(\mathcal{M}), \\ H_{\text{coord}} &= I_A = H_{|\mathcal{A}|}(\mathcal{A}), \end{aligned} \quad (24)$$

of a chemical, combinatorial and coordinational kind, respectively. This notation appears to be easier to memorize, too, and applies in the same way for the maximal, normal and total measures.

Fig. 2 gives an illustration of the three conceptually distinct contributions to a crystal structure's complexity emerging from the subdivision of a crystal structure composed of atoms, each having Z_i degrees of freedom, into Wyckoff positions, each contributing the pair (M_i, A_i) of degrees of freedom.

3. Extending Krivovichev complexity: bivariate case

While it is advantageous to have complexity measures accounting for the chemical, combinatorial and coordinational degrees of freedom separately, yielding a finer mode of analysis for conceptually different contributions to the total structural complexity, one might eventually prefer a single combined complexity measure, concise yet comprehensive, in one's toolbox.

In the following we will focus on the pairwise combination of complexity measures, since it will be shown that the combination of any number of complexity measures into a single one follows the same rules.

3.1. Fundamental properties of a combined Shannon entropy

Reflecting the fact that the entropy is an extensive property of a system, and with the individual contributions being treated on an equal footing, their combination should be additive in nature. Moreover, a combined measure based on individual Shannon entropies should be a Shannon entropy again, *i.e.* fulfilling all of its properties in general. In particular, the property of completeness [cf. equation (6)] of all the involved probability distributions, the individual ones as well as their combination, turns out to be decisive. However, the simple addition of entropies,

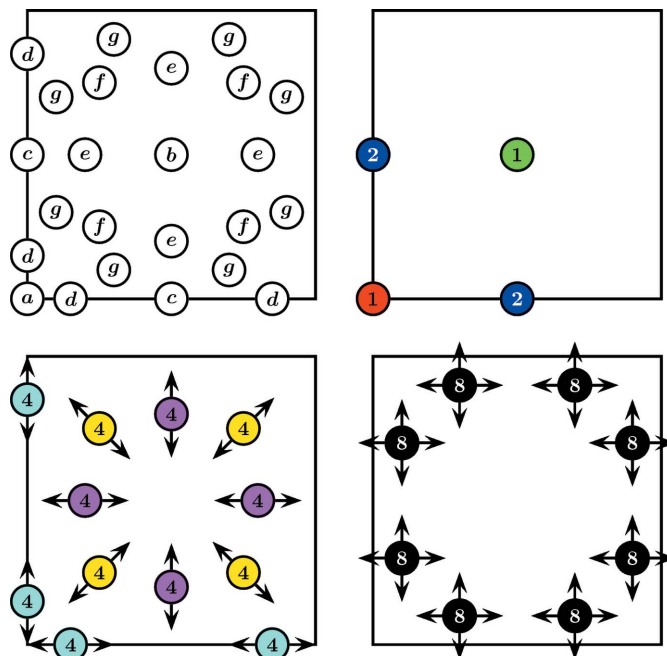


Figure 2

The Wyckoff positions in plane group $p4mm$ (No. 11), (top left) designated by their Wyckoff letters, and shown separately for positions with (top right) zero, (bottom left) one and (bottom right) two degrees of freedom (arities) A_i , as indicated by the presence of zero, one and two (orthogonal) pairs of arrows, respectively. Note that sites equivalent by symmetry can only move in unison. The corresponding multiplicities M_i are given by the number of visible sites of the same colour, as well as by the integers inside the circles representing a site. The possibilities of independent decoration of sites by atoms of atomic number Z_i are highlighted by giving each distinct Wyckoff position its own colour, red, green, blue, cyan, magenta, yellow and black, respectively. Sites equivalent by translation symmetry are not shown.

$$H_{|\mathcal{X}|}(\mathcal{X}) + H_{|\mathcal{Y}|}(\mathcal{Y}) = \sum_{i=1}^{|\mathcal{X}|} L\left(\frac{X_i}{X}\right) + \sum_{j=1}^{|\mathcal{Y}|} L\left(\frac{Y_j}{Y}\right), \quad (25)$$

fails to fulfil this condition by definition. (Similar issues caused by failing definitions of Shannon-like entropies arise in their application to the quantum-chemical analysis of atoms and molecules; see Flores-Gallegos, 2019.) The only proper way to combine a pair of individual probability distributions \mathcal{X} and \mathcal{Y} into a combined one, symbolized as $\mathcal{X} \uplus \mathcal{Y}$, such that the combined probabilities add up to unity, is given by

$$\mathcal{X} \uplus \mathcal{Y} = \frac{[X_i]_{i=1}^{|\mathcal{X}|} \uplus [Y_j]_{j=1}^{|\mathcal{Y}|}}{X + Y}. \quad (26)$$

Here, \uplus denotes a multiset sum in which the multiplicity of an element in the union multiset, $\mathcal{X} \uplus \mathcal{Y}$, is the sum of its multiplicities in the summand multisets, \mathcal{X} and \mathcal{Y} , respectively. For instance,

$$\frac{[4, 4, 4, 2, 2]}{16} \uplus \frac{[2, 1, 1]}{4} = \frac{[4, 4, 4, 2, 2, 2, 1, 1]}{20}. \quad (27)$$

Thus the resulting Shannon entropy, accounting for every degree of freedom in a proper manner, is given by

$$H_{|\mathcal{X}|+|\mathcal{Y}|}(\mathcal{X} \uplus \mathcal{Y}) = \sum_{i=1}^{|\mathcal{X}|} L\left(\frac{X_i}{X+Y}\right) + \sum_{j=1}^{|\mathcal{Y}|} L\left(\frac{Y_j}{X+Y}\right). \quad (28)$$

The reader is invited to compare this equation (28) with the previous equation (25) to spot their difference. In particular, one should note that, while the summands on the right-hand side of equation (25) are based on complete probability distributions/proper Shannon entropies whereas their sum by simple addition on the left-hand side of equation (25) is not, the opposite is true for equation (28).

Equation (28) covers a distinct form of additivity, known as strong additivity, in contrast to the plain additivity expressed in equation (25). At their core these two notions of additivity for the Shannon entropy (another one is subadditivity) are rooted in the question of the independence of the information. Additivity requires independence, *i.e.* the information content of the combination of independent subsystems is just the sum of their individual entropies. Strong additivity, by comparison, is conceptually more than that, the subsystems are not independent, and the total entropy contains properly weighted contributions from the conditional entropy of the individual subsystems in relation to the entropy of the system (see Appendix B for further formulas).

In a most remarkable way, both notions of additivity can be reconciled within a single mathematical expression for two probability distributions (subsystems), \mathcal{X} and \mathcal{Y} , namely as

$$H_{|\mathcal{X}|+|\mathcal{Y}|}(\mathcal{X} \uplus \mathcal{Y}) = H(X, Y) + \frac{X}{X+Y} \times H_{|\mathcal{X}|}(\mathcal{X}) + \frac{Y}{X+Y} \times H_{|\mathcal{Y}|}(\mathcal{Y}), \quad (29)$$

such that now both sides of the equation consist of proper Shannon entropies only. The individual entropies enter the equation in a weighted fashion, according to a subsystem's ratio within the system, while one additional entropy term, $H(X, Y)$, accounts for a contribution due to their combination, in particular depending on the subsystem's individual degrees of freedom, X and Y , respectively.

Let $\mathcal{X}^{(i)}$ denote the i th out of a set of S probability distributions (subsystems), each of order $|\mathcal{X}^{(i)}|$ and with $X^{(i)}$ contributing total degrees of freedom. Then the generalized strong additivity formula becomes

$$H_{\sum_{i=1}^S |\mathcal{X}^{(i)}|} \left(\biguplus_{i=1}^S \mathcal{X}^{(i)} \right) = H_S(X^{(1)}, \dots, X^{(S)}) + \sum_{i=1}^S \frac{X^{(i)} \times H_{|\mathcal{X}^{(i)}|}(\mathcal{X}^{(i)})}{X^{(1)} + \dots + X^{(S)}}. \quad (30)$$

In an axiomatic treatment of Shannon entropy it turns out that, together with continuity (small changes in the probabilities yield only small changes in the entropy) and symmetry (permutations of the probabilities leave the entropy invariant), strong additivity is, up to a multiplicative constant (normalization), the defining property of the Shannon entropy, distinguishing it uniquely from other entropy measures.

3.2. Compositional and configurational complexity

Using the strong additivity property of the Shannon entropy, we can now construct any combination of univariate entropy-based complexity measures. In particular, these are the bivariate compositional complexity,

$$I_{ZM} = H(Z, M) + \frac{Z \times H_{|\mathcal{Z}|}(\mathcal{Z})}{Z+M} + \frac{M \times H_{|\mathcal{M}|}(\mathcal{M})}{Z+M} = H(Z, M) + \frac{H_{\text{chem, total}} + H_{\text{comb, total}}}{Z+M}, \quad (31)$$

and the bivariate configurational complexity,

$$I_{MA} = H(M, A) + \frac{M \times H_{|\mathcal{M}|}(\mathcal{M})}{M+A} + \frac{A \times H_{|\mathcal{A}|}(\mathcal{A})}{M+A} = H(M, A) + \frac{H_{\text{comb, total}} + H_{\text{coor, total}}}{M+A}. \quad (32)$$

Here, the more explicit notations $H_{|\mathcal{Z}|+|\mathcal{M}|}(\mathcal{Z} \uplus \mathcal{M})$ and $H_{|\mathcal{M}|+|\mathcal{A}|}(\mathcal{M} \uplus \mathcal{A})$ have been abbreviated to I_{ZM} and I_{MA} , respectively, following the notation of Krivovichev. Apart from the aforementioned formulas based on the strong additivity property of the Shannon entropy, highlighting the univariate contributions, their calculation follows the general scheme presented in Fig. 1, including their maximal, normal and total variants.

The adjectives used for the bivariate complexities are chosen to best represent the combinations of the univariate complexities involved in their definition. The term 'compositional' refers to the fact that, in order to define a composition of a chemical compound, one has to specify both the atom types (chemical complexity) and their stoichiometric proportions (combinatorial complexity). In the same way, in order to specify a crystal structure as a purely geometric point set (configuration) in three-dimensional space, one has to specify the site multiplicities (combinatorial complexity) and the site arities (coordination complexity) of the Wyckoff positions.

In principle, there would be another combination of attributes possible, namely the one between the atomic numbers Z_i and the arities A_i , but it seems that this specific combination lacks an interpretation of its physical meaning, so it is left out of consideration.

3.3. Wyckoff multiplicities and arities

In this and the following section we focus on the configurational complexity. Any Wyckoff position is characterized by two integer quantities: its multiplicity (all distinct values occurring for reduced cells, coinciding with the possible orders of the point groups in three dimensions),

$$M_i \in \{1, 2, 3, 4, 6, 8, 12, 16, 24, 48\}, \quad (33)$$

and its spatial degree of freedom (arity),

$$A_i \in \{0, 1, 2, 3\}. \quad (34)$$

Now, the total number of spatial degrees of freedom A for a crystal structure consisting of M atoms is given by the number of free parameters A_i that one is able to specify for each of its N individual Wyckoff positions of site multiplicity M_i . In full

Table 2

Frequency distribution of 1731 distinct Wyckoff positions according to their multiplicities M_i (related to the reduced unit cell) and arities A_i .

Each pair (M_i, A_i) is represented by the first instance, starting from the highest space-group type number, of a matching Wyckoff sequence, stated in brackets, consisting of the space-group type number and the Wyckoff letter (related to the not-necessarily reduced unit cell) as given by Aroyo (2016). Except for one notable case, a pair is non-uniquely associated with a Wyckoff position, with the exceptional pair (1, 3) uniquely representing the Wyckoff position $1a$ in space-group type $P1$ (No. 1). Note that six potential combinations of values – (16, 0), (24, 0), (48, 0), (48, 1), (16, 2) and (48, 2) – do not have a Wyckoff position associated with them. The marginal and total sums are also given.

| M_i | A_i | | | | | | | | Σ |
|----------|-------|------------------|-----|------------------|-----|------------------|-----|------------------|----------|
| | 0 | | 1 | | 2 | | 3 | | |
| 1 | 167 | (229, <i>a</i>) | 34 | (183, <i>a</i>) | 3 | (8, <i>a</i>) | 1 | (1, <i>a</i>) | 205 |
| 2 | 255 | (227, <i>b</i>) | 198 | (191, <i>e</i>) | 23 | (46, <i>b</i>) | 8 | (9, <i>a</i>) | 484 |
| 3 | 39 | (229, <i>b</i>) | 24 | (189, <i>g</i>) | 5 | (174, <i>k</i>) | 4 | (146, <i>b</i>) | 72 |
| 4 | 106 | (229, <i>c</i>) | 245 | (217, <i>c</i>) | 59 | (121, <i>i</i>) | 45 | (82, <i>g</i>) | 455 |
| 6 | 34 | (229, <i>d</i>) | 72 | (229, <i>e</i>) | 18 | (190, <i>h</i>) | 22 | (174, <i>l</i>) | 146 |
| 8 | 11 | (230, <i>b</i>) | 74 | (229, <i>f</i>) | 29 | (141, <i>h</i>) | 68 | (122, <i>e</i>) | 182 |
| 12 | 5 | (230, <i>d</i>) | 55 | (229, <i>h</i>) | 16 | (217, <i>g</i>) | 27 | (199, <i>c</i>) | 103 |
| 16 | 0 | | 5 | (230, <i>e</i>) | 0 | | 20 | (142, <i>g</i>) | 25 |
| 24 | 0 | | 14 | (230, <i>g</i>) | 10 | (229, <i>k</i>) | 25 | (220, <i>e</i>) | 49 |
| 48 | 0 | | 0 | | 0 | | 10 | (230, <i>h</i>) | 10 |
| Σ | 617 | | 721 | | 163 | | 230 | | 1731 |

analogy with a mechanical system, any atomic configuration is properly determined only if all of these independent parameters are specified by its structure description. Table 2 gives an overview of the frequencies of occurrence for each pair (M_i, A_i) among the total number of 1731 distinct Wyckoff positions.

The listing shows that the multiplicities and arities are not independent attributes but exhibit a correlation, such that smaller multiplicities, say $M_i < 6$, are more often associated with fixed positions of arity $A_i = 0$, while larger multiplicities $M_i \geq 6$ are more often associated with general positions of arity $A_i = 3$. In order to quantify this statement, one can resort to the tool of contingency analysis (see Appendix C). Then one finds a value of $C_{\text{norm}} = 0.615$ for Cramér's normalized contingency coefficient (see Blaikie, 2003, p. 100), thus favouring association/non-independence of the attributes. This can also be seen as a formal justification for the use of the strong additivity in the calculation of combined Shannon entropies, in particular the calculation of the bivariate configurational complexity from the univariate combinatorial and coordinational complexities, respectively.

3.4. Non-equivalent crystal structures of identical Krivovichev complexity

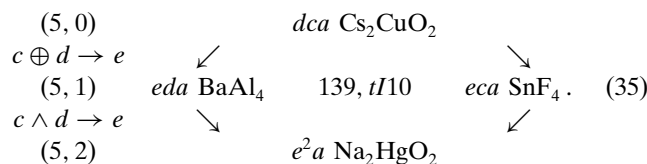
The reason for an inclusion of coordinate contributions to the structural complexity is made most obvious by considering those crystal structures which share the same Krivovichev complexities, yet differ in their degrees of freedom.

In order to have two crystal structures, in the following indexed as i and j , that share the same Krivovichev complexity, two conditions must be fulfilled: both crystal structures must share (i) the same number of atoms, $M_i = M_j$, and (ii) their multisets of fractional Wyckoff position multiplicities, $\mathcal{M}_i = \mathcal{M}_j$. The first condition is necessary but not sufficient, while the second condition implies the first, and thus can be used on its own.

Thus, we are interested in transformations of a crystal structure's Wyckoff sequence, such that what we will call its Wyckoff spectrum, *i.e.* \mathcal{M} , stays invariant. This can happen in different ways. A single Wyckoff position of given multiplicity and degree of freedom can be replaced in the Wyckoff sequence by (i) itself (identical case) or (ii) another Wyckoff position of the same multiplicity and degree of freedom, yet with a different Wyckoff letter, then $M_i = M_j$ and $A_i = A_j$ (isomorphous case), or (iii) another Wyckoff position of the same multiplicity but with a different degree of freedom (and Wyckoff letter), then $M_i = M_j$ but $A_i \neq A_j$ (non-isomorphous case).

Focusing on the non-isomorphous case for a single Wyckoff position replacement, this is possible in about half of all space-group types, namely in 100 out of 230 cases (43.5%), in particular in all of the space-group types listed in Table 3, together with all their possible Wyckoff positions as specified by their Wyckoff letters and differentiated for each multiplicity M_i according to the potential change in the degree of freedom A_i .

An example is given by the following scheme occurring for the space group type $I4/mmm$ (No. 139) for crystal structures of Pearson symbol $tI10$, *i.e.* all constituted of five atoms in the reduced cell, as expressed by the shorthand notation $(5, A)$, which, by the change in their Wyckoff sequence, gain additional degrees of freedom from $A = 0$ through $A = 1$ to $A = 2$ (\wedge = and, \oplus = exclusive or):



Note that all four structures share the same Wyckoff spectrum $\mathcal{M} = [2, 2, 1]/5$, thus sharing the same Krivovichev complexity (in bits per freedom)

$$H_{\text{comb}} = I_M = 2L(2/5) + L(1/5) = 1.522, \quad (36)$$

Table 3

Space-group types (listed by their numbers, SG No.) in which Wyckoff positions can be substituted keeping their multiplicity M_i fixed, while their arity A_i changes, thereby defining crystal structures with the same Krivovichev complexity despite their different numbers of geometric degrees of freedom.

| SG No. | M_i | $(M_i, 0)$ | $(M_i, 1)$ | $(M_i, 2)$ | SG No. | M_i | $(M_i, 0)$ | $(M_i, 1)$ | $(M_i, 2)$ |
|--------|-------|------------|------------|------------|--------|-------|------------|------------|------------|
| 10 | 2 | | $i-l$ | m, n | 125 | 4 | e, f | g, h | |
| 11 | 2 | $a-d$ | | e | | 8 | | $i-l$ | m |
| 12 | 2 | e, f | g, h | i | 126 | 4 | c, d | e | |
| 13 | 2 | $a-d$ | e, f | | | 8 | f | $g-j$ | |
| 15 | 2 | $a-d$ | e | | 128 | 4 | c, d | e | |
| 28 | 2 | | a, b | c | | 8 | | f, g | h |
| 35 | 2 | | c | d, e | 129 | 2 | a, b | c | |
| 38 | 2 | | c | d, e | | 4 | d, e | f | |
| 40 | 2 | | a | b | | 8 | | g, h | i, j |
| 42 | 2 | | b | c, d | 130 | 4 | a, b | c | |
| 46 | 2 | | a | b | | 8 | d | e, f | |
| 48 | 4 | e, f | $g-l$ | | 131 | 8 | | n | $o-q$ |
| 49 | 4 | | $i-p$ | q | 132 | 4 | e, f | $g-j$ | |
| 50 | 4 | e, f | $g-l$ | | | 8 | | $k-m$ | n, o |
| 51 | 2 | $a-d$ | e, f | | 133 | 8 | e | $f-j$ | |
| | 4 | | g, h | $i-k$ | 134 | 4 | $c-f$ | g | |
| 52 | 4 | a, b | c, d | | | 8 | | $h-l$ | m |
| 53 | 4 | | $e-g$ | h | 135 | 8 | | $e-g$ | h |
| 54 | 4 | a, b | c, d | | 136 | 4 | c, d | $e-g$ | |
| 55 | 4 | | e, f | g, h | | 8 | | h | i, j |
| 56 | 4 | a, b | c, d | | 137 | 8 | e | f | g |
| 57 | 4 | a, b | c | d | 138 | 4 | $a-d$ | e | |
| 58 | 4 | | e, f | g | 139 | 2 | c, d | e | |
| 59 | 4 | c, d | | e, f | | 4 | f | $g-j$ | |
| 60 | 4 | a, b | c | | | 8 | | k | $l-n$ |
| 62 | 4 | a, b | | c | 140 | 4 | e | $f-h$ | |
| 63 | 2 | a, b | c | | | 8 | | i, j | k, l |
| | 4 | d | e | f, g | 141 | 4 | c, d | e | |
| 64 | 4 | c | d, e | f | | 8 | | f, g | h |
| 65 | 2 | e, f | $g-l$ | | 142 | 8 | c | $d-f$ | |
| | 4 | | m | $n-q$ | 162 | 2 | c, d | e | |
| 66 | 4 | | $g-k$ | l | | 6 | | i, j | k |
| 67 | 2 | $a-f$ | g | | 163 | 6 | g | h | |
| | 4 | | $h-l$ | m, n | 164 | 6 | | g, h | i |
| 68 | 4 | c, d | $e-h$ | | 165 | 6 | e | f | |
| 69 | 2 | $c-f$ | $g-i$ | | 166 | 6 | | f, g | h |
| | 4 | | $j-l$ | $m-o$ | 167 | 6 | d | e | |
| 70 | 4 | c, d | $e-g$ | | 175 | 2 | c, d | e | |
| 71 | 4 | k | | $l-n$ | | 6 | | i | j, k |
| 72 | 4 | e | $f-i$ | j | 176 | 6 | g | | h |
| 73 | 4 | a, b | $c-e$ | | 177 | 2 | c, d | e | |
| 74 | 2 | $a-d$ | e | | 188 | 6 | | j | k |
| | 4 | | f, g | h, i | 189 | 2 | c, d | e | |
| 83 | 2 | e, f | g, h | | 190 | 6 | | g | h |
| | 4 | | i | j, k | 191 | 2 | c, d | e | |
| 84 | 4 | | $g-i$ | j | 192 | 4 | c, d | e | |
| 85 | 2 | a, b | c | | | 12 | | $i-k$ | l |
| | 4 | d, e | f | | 193 | 4 | c, d | e | |
| 86 | 4 | c, d | e, f | | | 6 | f | g | |
| 87 | 2 | c, d | e | | | 12 | | i | j, k |
| | 4 | f | g | h | 194 | 6 | g | h | |
| 88 | 4 | c, d | e | | | 12 | | i | j, k |
| 89 | 2 | e, f | g, h | | 202 | 6 | d | e | |
| 90 | 2 | a, b | c | | | 12 | | g | h |
| 97 | 2 | c, d | e | | 209 | 6 | d | e | |
| 98 | 4 | c | $d-f$ | | 211 | 6 | d | e | |
| 101 | 4 | | c | d | 215 | 12 | | h | i |
| 102 | 4 | | b | c | 217 | 6 | d | e | |
| 111 | 2 | e, f | g, h | | | 12 | | f | g |
| | 4 | | $i-m$ | n | 222 | 12 | d | e | |
| 113 | 2 | a, b | c | | 223 | 24 | | j | k |
| | 4 | | d | e | 224 | 12 | f | g | |
| 115 | 4 | | h, i | j, k | | 24 | | $h-j$ | k |
| 119 | 4 | | g, h | i | 225 | 6 | d | e | |
| 121 | 2 | c, d | e | | 226 | 24 | | h | i |
| | 4 | | $f-h$ | i | 229 | 6 | d | e | |
| 123 | 2 | e, f | g, h | | | 24 | | i | j, k |
| 124 | 4 | e, f | g, h | | | | | | |
| | 8 | | $i-l$ | m | | | | | |

but differ in their total degree of freedom A by one or two, thus exhibiting different coordinational complexities (in bits per freedom):

$$\begin{aligned} H_{\text{coord}} &= I_A(5, 0) = 0, \\ H_{\text{coord}} &= I_A(5, 1) = 0, \\ H_{\text{coord}} &= I_A(5, 2) = 1. \end{aligned} \quad (37)$$

With the individual contributions of the combinatorial and coordinational complexities known, one calculates the combined configurational complexity according to the aforementioned strong additivity formula of the Shannon entropy as

$$H_{\text{conf}} = I_{MA}(5, A) = H(5, A) + \frac{5 \times H(2, 2, 1) + A \times H_A}{5 + A}. \quad (38)$$

Plugging in the different values for A one obtains

$$\begin{aligned} H_{\text{conf}} &= I_{MA}(5, 0) = 0.000 + 1.522 + 0.000 = 1.522, \\ H_{\text{conf}} &= I_{MA}(5, 1) = 0.650 + 1.268 + 0.000 = 1.918, \\ H_{\text{conf}} &= I_{MA}(5, 2) = 0.863 + 1.087 + 0.286 = 2.236, \end{aligned} \quad (39)$$

with some of the intermediate terms, in fact, vanishing, and all final values given in bits per freedom.

Note how the configurational complexity $I_{MA}(5, A)$ rises, due to the non-vanishing exchange term (first term), even while the combinatorial complexity term (second term) drops and the coordinational complexity term (third term) stays zero, on changing A from zero to one. Note also how the configurational complexity rises, while the contribution of the combinatorial complexity to it gradually diminishes from 100% through 66.1% to 48.6%.

4. Strong additivity in crystallography

As it happens, the strong additivity property of the Shannon entropy transfers in a very natural way to various crystallographic contexts.

4.1. Strong additivity and structural hierarchies

In order to illustrate the rather abstract mathematical formulas of the previous paragraphs, we discuss the structure of a stuffed variant of the β -manganese structure type of Wyckoff sequence

$$212, d_{\text{Mo}}^1 c_{\text{Al}}^1 a_{\text{C}}^1 \quad (40)$$

as it occurs for the compound $cP24\text{-Mo}_3\text{Al}_2\text{C}$. Here, the Wyckoff sequence is stated in an extended form, including the atom types occupying a given site as a subscript to each Wyckoff letter.

Now, the probability distributions, which can be inferred from the extended Wyckoff sequence, and with entries listed in the order in which they occur within it, are given as

$$\begin{aligned} \mathcal{Z} &= [42, 13, 6]/61, \\ \mathcal{M} &= [12, 8, 4]/24, \\ \text{and } \mathcal{A} &= [1, 1]/2, \end{aligned} \quad (41)$$

respectively.

Note that, for the probability distribution of the arities \mathcal{A} , the expansibility property of the Shannon entropy has been used, since the fixed site of Wyckoff letter a does not contribute a degree of freedom. The special cases for crystal structures containing fixed sites, either with some or all $A_i = 0$, with $A = 0$ in addition for the latter case, have been explicitly considered by the case distinction stated in equation (2), thus preventing errors due to the non-definiteness of taking the logarithm of or performing a division by zero, respectively. Note that the coordinational complexity is also zero, $I_A = 0$, in those cases in which a crystal structure consists of a single site only. Then, $A_i = A$, and hence the logarithm of unity becomes zero.

As before, the focus will be on the configurational complexity, with the corresponding probability distributions being \mathcal{M} and \mathcal{A} , respectively. Let us recall the calculation of the configurational complexity [cf. equation (32)], now with the corresponding values of the probability distributions of our example being plugged in:

$$\begin{aligned} H(12, 8, 4, 1, 1) &= H(24, 2) + \frac{24}{26} H(12, 8, 4) + \frac{2}{26} H_2 \\ 1.815 &= 0.391 + 1.347 + 0.077. \end{aligned} \quad (42)$$

It can be seen that the structural complexity of the stuffed variant of the β -manganese structure type is 74.2% contributed by the combinatorial degrees of freedom of the structure, only 4.2% contributed by the coordinational degrees of freedom, and as much as 21.5% contributed by the term accounting for the combination of both subsystems' degrees of freedom into one system.

As was shown in equation (30), the strong additivity formula can be generalized to any number of subsystems in a straightforward manner by summing all weighted individual subsystem entropy terms together with one term for their combined entropy. Now, it is interesting to note that one can selectively calculate the combinatorial complexity of the stuffed β -manganese structure *in just the same manner*, namely as

$$\begin{aligned} H_{24} &= H(12, 8, 4) + \frac{12}{24} H_{12} + \frac{8}{24} H_8 + \frac{4}{24} H_4 \\ 4.585 &= 1.459 + 1.793 + 1.000 + 0.333. \end{aligned} \quad (43)$$

Note that the quantities H_{24} and $H_{24}(12, 8, 4)$ are identical to $I_{M, \text{max}}$ and I_M in the previous notation, which are thus found to be interrelated by the strong additivity property of the Shannon entropy. Here, the system consists of three Wyckoff position subsystems, of multiplicity twelve, eight and four, instead of the two subsystems of combinatorial and coordinational complexity as in equation (42) above.

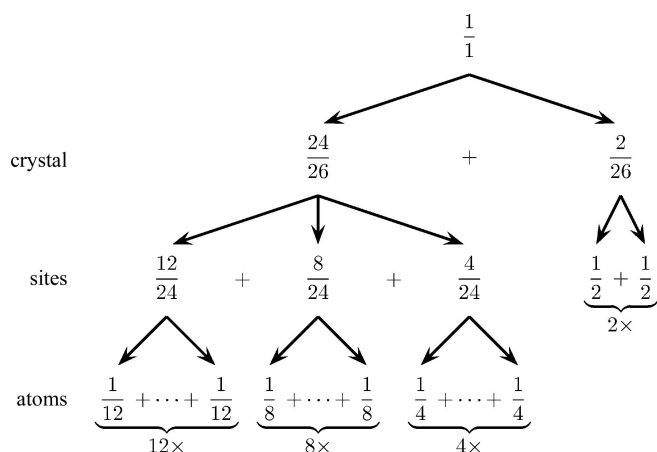


Figure 3

The taxonomic tree of complete probability distributions used in the calculation of the configurational and combinatorial entropies for the filled β -manganese crystal structure of Wyckoff sequence 212/213 *dca*. Note that while for each splitting of a system into subsystems the denominator of the ratios changes, their sum always adds to unity.

And again, the same equation (30) can be used even down to the level where the system is a single Wyckoff position of multiplicity M_i and the subsystems are single atoms:

$$H_{M_i} = H_{M_i} + M_i \times \frac{1}{M_i} H_1 = \log_2 M_i. \quad (44)$$

The latter case, however degenerate and trivial it appears from the viewpoint of calculation (ironically, almost like the construction of the set-theoretical von Neumann ordinals from the empty set, higher-level structural complexity seemingly arises out of the zero entropy terms H_1 associated with single-atom subsystems), shows nevertheless the hierarchical applicability of the Shannon entropy formula, integrating subsystems from several atoms to a single Wyckoff position, from several Wyckoff positions to a single-crystal structure, and from several individual contributions to one combined complexity measure.

The three levels of hierarchy reflected by the three equations above can be neatly presented within a taxonomic tree (Fig. 3). Note how, on each level of hierarchy and for each subdivision, the probabilities each add up to unity, as expected for a complete probability distribution, from which a Shannon entropy can be calculated. Note also how the subdivision process stops whenever a probability distribution with equidistributed probability values results.

4.2. Strong additivity and group–subgroup relations

As we have seen, the strong additivity property of the Shannon entropy can be used in the conceptual process of constructing a *single* crystal structure by integrating atoms into Wyckoff positions and Wyckoff positions into the final crystal structure. However, a similar scheme can be applied for the comparison of a *pair* of crystal structures exhibiting a group–subgroup relation. For this purpose we make use of another

Table 4

Classifying 15 503 distinct Wyckoff sequences (structure types) according to their combinatorial (I_M) and coordinational (I_A) complexities.

Note that M denotes the number of atoms in the reduced cell.

| $M \circ A$ | $I_M \circ I_A$ | No. of cases | % of cases |
|-------------|-----------------|--------------|------------|
| $M > A$ | $I_M > I_A$ | 9520 | 61.41 |
| $M = A$ | $I_M > I_A$ | 4 | 0.03 |
| $M < A$ | $I_M > I_A$ | 697 | 4.50 |
| $M > A$ | $I_M = I_A$ | 1482 | 9.56 |
| $M = A$ | $I_M = I_A$ | 496 | 3.20 |
| $M < A$ | $I_M = I_A$ | 466 | 3.01 |
| $M > A$ | $I_M < I_A$ | 2701 | 17.42 |
| $M = A$ | $I_M < I_A$ | 0 | 0.00 |
| $M < A$ | $I_M < I_A$ | 137 | 0.88 |
| Σ | | 15 503 | 100.01 |

algebraic property of the Shannon entropy, namely its recursivity, expressed as

$$H_N(p_1, p_2, p_3, \dots, p_N) = H_{N-1}(p_{12}, p_3, \dots, p_N) + p_{12} \times H(P_1, P_2). \quad (45)$$

Here, $p_{12} = p_1 + p_2$, with recursivity naturally holding for any choice of pair $p_{ij} > 0$. In fact, recursivity just describes a special case of strong additivity (Baez *et al.*, 2011).

Now, think of a pair of crystal structures related by a group–subgroup transformation involving the splitting of a Wyckoff position, of multiplicity m_{12} , into a pair of Wyckoff positions, of multiplicities m_1 and m_2 . Recalling that $p_i = M_i/M$, equation (45) acquires a corresponding crystallographic meaning, as relating the Shannon entropy of the crystal structure higher in symmetry, $H_{N-1}(p_{12}, p_3, \dots, p_N)$, to the one lower in symmetry, $H_N(p_1, p_2, p_3, \dots, p_N)$. Symmetry reduction naturally increases structural complexity, and by applying equation (45) we can quantify the exact amount of this complexity increase as a result of symmetry reduction, namely $p_{12} \times H(P_1, P_2)$, which is, of course, just the entropy difference between the two crystal structures.

4.3. Strong additivity and crystal structure classification

One advantage inherent in equation (30) is the flexibility of defining new complexity measures on subsystems (*e.g.* for partial structures of crystal structures), as well as the possibility of splitting one complexity value into its individual contributions, which could be called complexity decomposition analysis (CDA). This lends itself to a novel method of crystal structure classification.

For instance, structures in which the combinatorial complexity turns out to be more important than the coordinational complexity, *i.e.* for which $I_M > I_A$ holds true (combinatorial-complex case), will form their own separate class in this scheme, as will be the case for those structures for which the opposite is true, *i.e.* for which $I_M < I_A$, or those where both chemical and coordinate complexity are balanced, *i.e.* for which $I_M = I_A$. Since it can happen that the relation between the number of atoms M and the number of degrees of freedom A does not follow the same trend as the corresponding

complexities, one might introduce an even finer classification scheme according to the relations $M > A$, $M = A$ and $M < A$, respectively. Table 4 contains the statistics of structures according to this finer classification scheme, again based on the 15 503 distinct Wyckoff sequences present in the PCD.

Following this classification scheme, there are a total of 10 221 (65.9%) structures for which $I_M > I_A$, only 2838 (18.3%) represent the opposite trend of $I_M < I_A$ and about the same number, 2444 (15.8%), are balanced ($I_M = I_A$).

In general, structures for which $I_M < I_A$ (coordination-complex case) can occur mostly in those space-group types in which the multiplicity of the Wyckoff positions in the reduced cell is smaller than its number of degrees of freedom, $M_i < A_i$. Of 1731 Wyckoff positions in total, the overwhelming majority, namely 1658 (95.8%), are characterized by $M_i > A_i$, while for 61 (3.5%) positions $M_i = A_i$ and for only 12 (0.7%) positions $M_i < A_i$. These 12 positions occur as the general positions of the triclinic space-group types, as well as the general and some special positions of a few monoclinic space-group types, listed in the following by their space-group number/Wyckoff letter/ M_i/A_i : 1/*a*/1/3, 2/*i*/2/3, 3/*e*/2/3, 4/*a*/2/3, 5/*c*/2/3, 6/*a*/1/2, 6/*b*/1/2, 6/*c*/2/3, 7/*a*/2/3, 8/*a*/1/2, 8/*b*/2/3 and 9/*a*/2/3. In four of these space-group types the only existing position is the general position. Thus, it happens that many coordination-complex structures are solely composed of multiple occurrences of the general position. For instance, the structure type of the complex phosphate *mP*882-Mo₂P₄O₁₅, with $M = 882$, $A = 1323$ and Wyckoff sequence 7, *a*⁴⁴¹, belongs to this kind of structure, since the general Wyckoff position, and the only one existing in space-group type *Pc*, is the one with the symbol 2*a* ($M_i < A_i$) (441 being almost the highest power occurring in any Wyckoff sequence, with the record of 461 associated with the Wyckoff sequence 11, *f*⁴⁶¹*e*²⁶ which occurs for the structure type *mP*1896-Na₁₅Fe₃Co₁₆[Mo₁₇₆O₅₂₈H₃(H₂O)₈₀]Cl₂₇·450H₂O with $I_M > I_A$, since the general position has symbol 4*f*).

Another reason for the observed frequency of occurrence of combinatorial-complex and coordination-complex structures is given by the composition observed for the sets of multiplicities and arities. With

$$\langle M_i \rangle = \frac{1 + 2 + 3 + 4 + 6 + 8 + 12 + 16 + 24 + 48}{10} = 12.4, \quad (46)$$

the average multiplicity, with respect to all possible values, is much greater than the average degree of freedom,

$$\langle A_i \rangle = \frac{0 + 1 + 2 + 3}{4} = 1.5. \quad (47)$$

Taking the average with respect to the Wyckoff positions occurring in the individual space-group types (subscript SGT), the variation is greater

$$1.0 \text{ (No. 1)} < \langle M_i \rangle_{\text{SGT}} < 19.0 \text{ (No. 230)}, \quad (48)$$

$$0.3 \text{ (No. 2)} < \langle A_i \rangle_{\text{SGT}} < 3.0, \quad (49)$$

Table 5

Frequency distribution of 15 503 distinct Wyckoff sequences according to their combinatorial ($I_{M, \text{total}}$) and coordinational ($I_{A, \text{total}}$) complexities, binned into intervals of powers of ten.

Here, the notation $\langle i, j \rangle$ is shorthand for the interval $[10^i, 10^j)$. The marginal and total sums are also given.

| $I_{M, \text{total}}$ | $I_{A, \text{total}}$ | | | | | | Σ |
|------------------------|-----------------------|----------|------------------------|------------------------|------------------------|------------------------|----------|
| | $[0, 1)$ | $(0, 1)$ | $\langle 1, 2 \rangle$ | $\langle 2, 3 \rangle$ | $\langle 3, 4 \rangle$ | $\langle 4, 5 \rangle$ | |
| $[0, 1)$ | 46 | 0 | 0 | 0 | 0 | 0 | 46 |
| $(0, 1)$ | 253 | 168 | 8 | 0 | 0 | 0 | 429 |
| $\langle 1, 2 \rangle$ | 321 | 1702 | 3637 | 99 | 0 | 0 | 5759 |
| $\langle 2, 3 \rangle$ | 3 | 73 | 3808 | 4221 | 107 | 0 | 8212 |
| $\langle 3, 4 \rangle$ | 0 | 0 | 31 | 521 | 489 | 3 | 1044 |
| $\langle 4, 5 \rangle$ | 0 | 0 | 0 | 0 | 11 | 2 | 13 |
| Σ | 623 | 1943 | 7484 | 4841 | 607 | 5 | 15 503 |

with $\langle A_i \rangle_{\text{max}} = 3.0$ achieved for space-group types of number 1, 4, 7, 9, 19, 29, 33, 76, 78, 144, 145, 169 and 170. The averaged averages, however, still show the same trend, albeit more diminished,

$$\langle \langle M_i \rangle_{\text{SGT}} \rangle = 4.8, \quad (50)$$

$$\langle \langle A_i \rangle_{\text{SGT}} \rangle = 1.3, \quad (51)$$

of $\langle \langle M_i \rangle_{\text{SGT}} \rangle > \langle \langle A_i \rangle_{\text{SGT}} \rangle$. Thus, assuming everything else to be equal, combinatorial-complex structures should occur roughly four times more frequently than coordination-complex structures.

A more detailed picture can be obtained by looking at the distribution of crystal structures according to their combinatorial and coordinate complexity values. Table 5 contains the absolute frequencies of occurrence for 15 503 distinct Wyckoff sequences distributed according to their $I_{M, \text{total}}$ and $I_{A, \text{total}}$ values into 36 bins, with equidistributed interval limits chosen with respect to powers of ten, and according to a logarithmic subdivision of values ranging over six orders of magnitude.

As was done above for the individual Wyckoff positions and the association of their attributes' multiplicity and arity, a contingency analysis reveals a Cramér's normalized contingency coefficient of $C_{\text{norm}} = 0.799$, again favouring association of the corresponding complexity measures, $I_{M, \text{total}}$ and $I_{A, \text{total}}$, respectively.

5. Extending Krivovichev complexity: trivariate case

At this point in our analysis it makes sense to review the unifying notation for all the complexity measures we have presented. For this purpose we collect the aforementioned measures conceptually into the 'six C's of complexity' (Fig. 4).

The scheme presented in Fig. 4 covers all the information of an ordered (no mixed-occupancy sites are present and all occupancy parameters are unity), static (no atomic displacement parameters are considered) and affine (the lattice metrics is ignored) crystal structure, in which each atomic site is specified by a triple of integers (Z_i, M_i, A_i), given by the

Table 6

Values of the chemical, combinatorial, coordinational, compositional, configurational and crystallographical complexities for the crystal structure of *cP24-Mo₃Al₂C*.

The units are bits per freedom for the conventional (non-subscripted) and maximal Shannon entropies (Krivovichev complexities), and bits per unit cell for the total entropy, while the normal one is a dimensionless quantity ranging between zero and unity. The Shannon entropy of complexity designator *xxxx* (chem, comb, coor, comp, conf, crys) corresponds to the Krivovichev complexity of complexity attribute *X* (*Z*, *M*, *A*, *ZM*, *MA*, *ZMA*).

| Shannon entropy | Krivovichev complexity | chem <i>Z</i> | comb <i>M</i> | coor <i>A</i> | comp <i>ZM</i> | conf <i>MA</i> | crys <i>ZMA</i> |
|--------------------------|------------------------|------------------|------------------|------------------|-------------------|-------------------|--------------------|
| H_{xxxx} | I_X | 1.175 | 1.459 | 1.000 | 2.114 | 1.815 | 2.246 |
| $H_{xxxx, \max}$ | $I_{X, \max}$ | 5.931 | 4.585 | 1.000 | 6.409 | 4.700 | 6.443 |
| $H_{xxxx, \text{norm}}$ | $I_{X, \text{norm}}$ | 0.198 | 0.318 | 1.000 | 0.330 | 0.386 | 0.349 |
| $H_{xxxx, \text{total}}$ | $I_{X, \text{total}}$ | 71.682 | 35.020 | 2.000 | 179.686 | 47.192 | 195.424 |

atomic number of the chemical element occupying the site and its Wyckoff multiplicity and arity.

The corresponding equations, here just given for the total entropies (the other entropies being defined according to the scheme presented in Fig. 1), are:

$$H_{\text{chem, total}} = Z \times H_{|Z|}(Z), \quad (52)$$

$$H_{\text{comb, total}} = M \times H_{|M|}(M), \quad (53)$$

$$H_{\text{coor, total}} = A \times H_{|A|}(A), \quad (54)$$

$$H_{\text{comp, total}} = (Z + M) \times H(Z, M) + H_{\text{chem, total}} + H_{\text{comb, total}}, \quad (55)$$

$$H_{\text{conf, total}} = (M + A) \times H(M, A) + H_{\text{comb, total}} + H_{\text{coor, total}}, \quad (56)$$

$$H_{\text{crys, total}} = (Z + M + A) \times H(Z, M, A) + H_{\text{chem, total}} + H_{\text{comb, total}} + H_{\text{coor, total}}. \quad (57)$$

Note that $H_{\text{crys, total}}$ can be defined alternatively as

$$H_{\text{crys, total}} = (Z + M + A) \times H(Z + M, A) + H_{\text{comp, total}} + H_{\text{coor, total}} \quad (58)$$

$$= (Z + M + A) \times H(Z, M + A) + H_{\text{chem, total}} + H_{\text{conf, total}}, \quad (59)$$

taking into account the compositional and configurational entropies, as indicated in the splitting shown in Fig. 4, respectively. Indeed, the transformation between these formulas is governed by the recursivity of the Shannon entropy:

$$H(Z, M, A) = H(Z + M, A) + \frac{Z + M}{Z + M + A} \times H(Z, M) \quad (60)$$

$$= H(Z, M + A) + \frac{M + A}{Z + M + A} \times H(M, A). \quad (61)$$

As emphasized before, this facilitates a more refined analysis of complex crystal structures, since it becomes possible to differentiate between crystal structures in a classification scheme according to their distinct complexities. For instance, for the aforementioned example given by the compound *cP24-Mo₃Al₂C*, the calculated complexities using the general scheme of Fig. 1 and equations (52) to (57) are listed in Table 6 as a reference. Here, the full set of conventional (non-subscripted), maximal, normal and total Shannon entropies (Krivovichev complexities) are given according to their subscripted complexity designator *xxxx* (= chem, comb, coor, comp, conf or crys in Shannon entropy notation) or attribute *X* (= *Z*, *M*, *A*, *ZM*, *MA* or *ZMA* in Krivovichev complexity notation). The reason for using two notations here is to invoke a consistent notation for the novel extended complexity measures, highlighting their character as Shannon entropies, while keeping a concordance with the complexity measure of Krivovichev, $I_G = I_M = H_{\text{comb}}$, upon which the extended measures are conceptually based.

$$\text{crystallographic } (Z, M, A) = \overbrace{\text{chemical } (Z) + \text{combinatorial } (M)}^{\text{compositional } (Z, M)} + \underbrace{\text{coordinational } (A)}_{\text{configurational } (M, A)}$$

Figure 4

The six C's of structural complexity, with their corresponding complexity attributes *X* stated in brackets (cf. Fig. 1). The complexity designators *xxxx* [cf. equations (52) to (57)] used as subscripts for the respective Shannon entropies are given by the first four letters of each adjective. Since each complexity can be associated with a tetrad of conventional, maximal, normal and total Shannon entropies, there is ultimately a full set of 24 distinguishable Shannon entropies (Krivovichev complexities) characterizing any crystal structure (cf. Table 6).

6. Conclusions

We have developed extensions to the complexity measures of Krivovichev, fully exploiting the information encoded in the extended, *i.e.* atom-type augmented, Wyckoff sequence of a crystal structure. As such they represent quantitative crystal structure descriptors (QCSDs; Hornfeck, 2012), such as those widely used in the chemoinformatics literature for molecules. These QCSDs can be correlated with other descriptors, such as selected physical properties of a compound, thereby facilitating a finer than existing quantitative classification scheme for crystal structures.

APPENDIX A

Algebraic transformations

In the following we derive equation (12) for the complexity measure $I_{G, \text{total}}$. We start with the definition of I_G ,

$$I_G = - \sum_{i=1}^N p_i \log_2 p_i = \sum_{i=1}^N p_i \log_2 \frac{1}{p_i}, \quad (62)$$

which after substituting the probabilities becomes

$$I_G = \sum_{i=1}^N \frac{M_i}{M} \log_2 \frac{M}{M_i}. \quad (63)$$

With $I_{G, \text{total}} = M \times I_G$ we have

$$I_{G, \text{total}} = \sum_{i=1}^N M_i \log_2 \frac{M}{M_i} \quad (64)$$

$$= \sum_{i=1}^N (M_i \log_2 M - M_i \log_2 M_i) \quad (65)$$

$$= \sum_{i=1}^N M_i \log_2 M - \sum_{i=1}^N M_i \log_2 M_i \quad (66)$$

$$= \log_2 M \sum_{i=1}^N M_i + \sum_{i=1}^N L(M_i) \quad (67)$$

$$= M \log_2 M + \sum_{i=1}^N L(M_i). \quad (68)$$

This last formula can be used to relate I_G and $I_{G, \text{max}}$:

$$I_G = I_{G, \text{max}} + \frac{1}{M} \sum_{i=1}^N L(M_i), \quad (69)$$

$$I_{G, \text{max}} - I_G = -\frac{1}{M} \sum_{i=1}^N L(M_i). \quad (70)$$

Since all $M_i \geq 1$ and thus all $L(M_i) \leq 0$ the final result is always non-negative. Alternatively, one can also write

$$I_{G, \text{max}} = \sum_{i=1}^N L(M_i/M) - \sum_{i=1}^N L(M_i)/M. \quad (71)$$

APPENDIX B

Notions of additivity

In accounting for the summation of Shannon entropies, one has to consider more than one notion of additivity. Indeed, the many characteristic algebraic properties of the Shannon entropy can be understood axiomatically (Aczél & Daróczy, 1975; Taneja, 2001; Csiszár, 2008) by asking, what kinds of properties should be natural for a proper measure of information content (Aczél *et al.*, 1974). It turns out that three properties concern additivity. For the first and second of these, let

$$\mathcal{P} = [p_1, p_2, \dots, p_U] \quad (72)$$

and

$$\mathcal{Q} = [q_1, q_2, \dots, q_V] \quad (73)$$

and

$$\begin{aligned} \mathcal{R} = & [p_1 q_1, p_1 q_2, \dots, p_1 q_V, \dots, \\ & p_2 q_1, p_2 q_2, \dots, p_2 q_V, \dots, \\ & p_U q_1, p_U q_2, \dots, p_U q_V]. \end{aligned} \quad (74)$$

The subadditivity is then expressed as (Aczél & Daróczy, 1975, ch. 1, p. 30, equation 1.2.7)

$$H_{U \cdot V}(\mathcal{R}) \leq H_U(\mathcal{P}) + H_V(\mathcal{Q}), \quad (75)$$

with additivity being the special case for equality. As before, $U = |\mathcal{P}|$ and $V = |\mathcal{Q}|$. For the third one, let

$$\mathcal{P} = [p_1, p_2, \dots, p_U] \quad (76)$$

and

$$\mathcal{Q}_i = [q_{i1}, q_{i2}, \dots, q_{iV(i)}] \quad (77)$$

and

$$\begin{aligned} \mathcal{R} = & [p_1 q_{11}, p_1 q_{12}, \dots, p_1 q_{1V(1)}, \dots, \\ & p_2 q_{21}, p_2 q_{22}, \dots, p_2 q_{2V(2)}, \dots, \\ & p_U q_{U1}, p_U q_{U2}, \dots, p_U q_{UV(U)}]. \end{aligned} \quad (78)$$

and strong additivity is then expressed as (Aczél & Daróczy, 1975, ch. 1, p. 30, equation 1.2.6)

$$H_{V(1)+V(2)+\dots+V(U)}(\mathcal{R}) = H_U(\mathcal{P}) + \sum_{i=1}^U p_i H_{V(i)}(\mathcal{Q}_i). \quad (79)$$

Note that $U = |\mathcal{P}|$, and the $V(i) = |\mathcal{Q}_i|$ are usually distinct from each other.

Thus, in a general axiomatic context this can be summarized as (Csiszár, 2008)

$$H(X, Y) \leq H(X) + H(Y), \quad (\text{subadditivity}) \quad (80)$$

$$H(\mathcal{P} \times \mathcal{Q}) = H(\mathcal{P}) + H(\mathcal{Q}), \quad (\text{additivity}) \quad (81)$$

$$H(X, Y) = H(X) + H(Y|X), \quad (\text{strong additivity}) \quad (82)$$

in which \mathcal{P} and \mathcal{Q} represent discrete probability distributions, X and Y represent distributions of random variables, $H(X, Y)$ denotes a joint entropy and $H(Y|X)$ denotes a conditional entropy.

APPENDIX C

Contingency analysis

Given two categorical variables X and Y , each with a certain number I and J of associated attributes x_1, \dots, x_I and y_1, \dots, y_J , respectively, one can construct a contingency matrix (also known as contingency table)

$$\begin{pmatrix} h_{11} & h_{12} & \dots & h_{1J} \\ h_{21} & h_{22} & \dots & h_{2J} \\ \vdots & \vdots & h_{ij} & \vdots \\ h_{I1} & h_{I2} & \dots & h_{IJ} \end{pmatrix} \quad (83)$$

whose entries h_{ij} represent the observed frequencies of occurrence for any given pairwise combination (x_i, y_j) of the attributes. Let

$$h_{\bullet j} = \sum_{i=1}^I h_{ij} \quad \text{and} \quad h_{i\bullet} = \sum_{j=1}^J h_{ij} \quad (84)$$

be the row and column marginal totals, respectively, and

$$h_{\bullet\bullet} = \sum_{i=1}^I \sum_{j=1}^J h_{ij} \quad (85)$$

be their grand total. Then Pearson's quadratic contingency (χ squared coefficient) is given as

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(h_{ij} - h)^2}{h}, \quad (86)$$

in which

$$h = \frac{h_{\bullet j} h_{i\bullet}}{h_{\bullet\bullet}}. \quad (87)$$

For a 2×2 contingency matrix, χ^2 varies from the value 1 (complete association) through 0 (no association) to -1 (complete inverse association), yet for the general case the range of values is different. In order to make the contingency measure comparable between contingency matrices of different general dimensions, one can use Cramér's contingency coefficient

$$C = \left(\frac{\chi^2}{\chi^2 + h_{\bullet\bullet}} \right)^{1/2} \quad (88)$$

in its normalized variant $C_{\text{norm}} = C/K$ with the normalization constant

$$K = \left(\frac{I-1}{I} \times \frac{J-1}{J} \right)^{1/4}, \quad (89)$$

depending on the dimensions of the contingency matrix. The obtained measure C_{norm} then takes values from zero to unity. A zero value is associated with full independence, as obtained by a contingency table with equidistributed frequency values, while a value of unity represents full association, as obtained for any permutation matrix including the identity matrix.

Thus, by aggregating and interpreting the observed frequencies of occurrence in this way, it is possible to assess the degree of statistical association/independence between the attributes.

Acknowledgements

The author thanks Morgane Poupon and Yaşar Krysiak for helpful discussions, and Michal Dušek for endorsement.

Funding information

This work was supported by the Czech Science Foundation through research grant No. 18-10438S, and by Project

No. LO1603 under the Ministry of Education, Youth and Sports National Sustainability Programme I of the Czech Republic.

References

- Aczél, J. & Daróczy, Z. (1975). *On Measures of Information and Their Characterizations. Mathematics in Science and Engineering*, Vol. 115. New York: Academic Press.
- Aczél, J., Forte, B. & Ng, C. T. (1974). *Adv. Appl. Probab.* **6**, 131–146.
- Aroyo, M. I. (2016). *International Tables for Crystallography*, Vol. A, *Space-group symmetry*. Chichester: Wiley.
- Baez, J. C., Fritz, T. & Leinster, T. (2011). *Entropy*, **13**, 1945–1957.
- Bertz, S. H. (1981). *J. Am. Chem. Soc.* **103**, 3599–3601.
- Bertz, S. H. (1983). *Bull. Math. Biol.* **45**, 849–855.
- Blaikie, N. (2003). *Analyzing Quantitative Data – From Description to Explanation*. London: SAGE Publications.
- Csiszár, I. (2008). *Entropy* **10**, 261–273.
- Dshemuchadse, J. & Steurer, W. (2015). *Inorg. Chem.* **54**, 1120–1128.
- Flores-Gallegos, N. (2019). *Chem. Phys. Lett.* **720**, 1–6.
- Hornfeck, W. (2012). *Acta Cryst.* **A68**, 167–180.
- Hornfeck, W. & Hoch, C. (2015). *Acta Cryst.* **B71**, 752–767.
- Krivovichev, S. (2012a). *Acta Cryst.* **A68**, 393–398.
- Krivovichev, S. V. (2012b). *Struct. Chem.* **23**, 1045–1052.
- Krivovichev, S. V. (2013a). *Microporous Mesoporous Mater.* **171**, 223–229.
- Krivovichev, S. V. (2013b). *Miner. Mag.* **77**, 275–326.
- Krivovichev, S. V. (2014a). *Miner. Mag.* **78**, 415–435.
- Krivovichev, S. V. (2014b). *Angew. Chem. Int. Ed.* **53**, 654–661.
- Krivovichev, S. V. (2016). *Acta Cryst.* **B72**, 274–276.
- Krivovichev, S. V. (2017). *Crystallogr. Rev.* **23**, 2–71.
- Krivovichev, S. V., Hawthorne, F. C. & Williams, P. A. (2017). *Struct. Chem.* **28**, 153–159.
- Krivovichev, S. V. & Krivovichev, V. G. (2020). *Acta Cryst.* **A76**, 429–431.
- Krivovichev, S. V., Krivovichev, V. G. & Hazen, R. M. (2018). *Eur. J. Mineral.* **30**, 231–236.
- Mackay, A. L. (1984). *Croat. Chem. Acta*, **57**, 725–736.
- Mackay, A. L. (2001). *Crystallogr. Rep.* **46**, 524–526.
- Parthé, E., Cenxual, K. & Gladyshevskii, R. E. (1993). *J. Alloys Compd.* **197**, 291–301.
- Parthé, E. & Gelato, L. M. (1984). *Acta Cryst.* **A40**, 169–183.
- Pettifor, D. G. (1984). *Solid State Commun.* **51**, 31–34.
- Pettifor, D. G. (1986). *J. Phys. C Solid State Phys.* **19**, 285–313.
- Rashevsky, N. (1955). *Bull. Math. Biophys.* **17**, 229–235.
- Shannon, C. E. (1948a). *Bell Syst. Tech. J.* **27**, 379–423.
- Shannon, C. E. (1948b). *Bell Syst. Tech. J.* **27**, 623–656.
- Siidra, O. I., Zenko, D. S. & Krivovichev, S. V. (2014). *Am. Mineral.* **99**, 817–823.
- Steurer, W. & Dshemuchadse, J. (2016). *Intermetallics – Structures, Properties and Statistics. IUCr Monographs on Crystallography*, No. 26. Oxford University Press.
- Tambornino, F., Sappl, J. & Hoch, C. (2015). *J. Alloys Compd.* **618**, 326–335.
- Taneja, I. J. (2001). *Generalized Information Measures and Their Applications*. Ebook, <http://www.mtm.ufsc.br/~taneja/book/book.html>. Retrieved 20/06/2019.
- Villars, P. & Cenxual, K. (2013). *Pearson's Crystal Data Crystal Structure Database for Inorganic Compounds*. ASM International, Materials Park, Ohio, USA. <http://www.crystalimpact.com/pcd/Default.htm>.
- Yodogawa, E. (1982). *Percept. Psychophys.* **32**, 230–240.